# CDC COVID-19 Case Surveillance Restricted Access Detailed Data
# COVID-19 Case Surveillance Data Access, Summary, Guidance, and Limitations

***CDC COVID-19 Emergency Response, October 2020***

***U.S. Centers for Disease Control and Prevention***

*Suggested Citation: Centers for Disease Control and Prevention, COVID-19 Response. COVID-19 Case Surveillance Restricted Data Access, Summary, and Limitations (version date: October 31, 2020).*

## Purpose

The purpose of this document is to facilitate proper access, analysis, and interpretation of the novel coronavirus (COVID-19) restricted case surveillance data. The document summarizes important information on the data access process and describes limitations of the case surveillance data.

## Introduction

The COVID-19 case surveillance system database includes individual-level data reported to U.S. states and autonomous reporting entities, including New York City and the District of Columbia (D.C.), as well as U.S. territories and states. On April 5, 2020, COVID-19 was added to the Nationally Notifiable Condition List and classified as "immediately notifiable, urgent (within 24 hours)" by a Council of State and Territorial Epidemiologists (CSTE) Interim Position Statement (Interim-20-ID-01). CSTE updated the position statement on August 5, 2020 to clarify the interpretation of antigen detection tests and serologic test results within the case classification. The statement also recommended that all states and territories enact laws to make COVID-19 reportable in their jurisdiction, and that jurisdictions conducting surveillance should submit case notifications to CDC. COVID-19 case surveillance data are collected by jurisdictions and shared voluntarily with CDC. For more information, https://wwwn.cdc.gov/nndss/conditions/coronavirus-disease-2019-covid-19/case-definition/2020/08/05/

The deidentified data in the restricted access dataset include demographic characteristics (including state and county), exposure history, disease severity indicators and outcomes, clinical data, laboratory diagnostic test results, and comorbidities. All data elements can be found on the COVID-19 case report form located at www.cdc.gov/coronavirus/2019-ncov/downloads/pui-form.pdf.

## Restricted Data Access Process

The Case Surveillance Task Force and Surveillance Review and Response Group (SRRG) within CDC's COVID-19 Emergency Response provide stewardship for datasets that support the public health community's access to COVID-19 data while protecting patient privacy. Data are made available for limited use upon completion of the registration information and data use restrictions agreement (RIDURA).

Next steps include completing the Restricted Access Data Use Restrictions Agreement (RIDURA) and it should be forwarded to Ask SRRG at eocevent394@cdc.gov. The request will be reviewed and if approved, expect to receive an email providing the GitHub requirements and instructions. If more information is required or the request is not approved, expect to receive email correspondence from ASK SRRG (eocvent394@cdc.gov).

## Restricted Data Specifications

A restricted access, detailed version of line-listed dataset of all COVID-19 cases reported to CDC is available. The dataset is to be made available for limited use upon completion of the RIDURA. COVID-19 data may differ substantially in the variables reported and in completeness by state. Some data are suppressed to protect individual privacy by coding as *NA* (see Data Suppression below). The ***restricted access*** data set includes the following variables:

- Initial case report date to CDC
- Date of first positive specimen collection
- Symptom onset date, if symptomatic
- Case status
- Sex
- Age group (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+ years)
- Race and ethnicity (combined)
- State of residence
- County of residence
- Healthcare worker status
- Pneumonia present
- Acute respiratory distress syndrome (ARDS) present
- Abnormal chest x-ray (CXR) present
- Hospitalization status
- ICU admission status
- Mechanical ventilation (MV)/intubation status
- Death status
- Presence of each of the following symptoms: fever, subjective fever, chills, myalgia, rhinorrhea, sore throat, cough, shortness of breath, nausea/vomiting, headache, abdominal pain, diarrhea
- Presence of underlying comorbidity or disease

## Public Data Specifications

A public version of line-listed dataset of all COVID-19 cases reported to CDC is available at **data.cdc.gov**. Completion of the RIDURA is not required. COVID-19 data may differ substantially in the variables reported and in completeness by state. Some data are suppressed to protect patient privacy by coding as *NA* (see Data Suppression below). The ***public use*** data set includes the following variables:

- Initial case report date to CDC
- Date of first positive specimen collection
- Symptom onset date, if symptomatic
- Case status
- Sex
- Age group (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+ years)
- Race and ethnicity (combined)
- Hospitalization status
- ICU admission status
- Mechanical ventilation (MV)/intubation status
- Death status
- Presence of underlying comorbidity or disease

## Case Data Standardization

COVID-19 case reports have been routinely submitted using standardized case reporting forms. On April 5, 2020, CSTE released an Interim Position Statement with national surveillance case definitions for COVID-19 included. CSTE updated the position statement on August 5, 2020 to clarify the interpretation of antigen detection tests and serologic test results within the case classification. The statement also recommended that all states and territories enact laws to make COVID-19 reportable in their jurisdiction, and that jurisdictions conducting surveillance should submit case notifications to CDC. COVID-19 case surveillance data are collected by jurisdictions and shared voluntarily with CDC. Current versions of these case definitions are available here: https://wwwn.cdc.gov/nndss/conditions/coronavirus-disease-2019-covid-19/. All cases reported on or after were requested to be reported by public health departments to CDC using the standardized case definitions for lab-confirmed or probable cases. On May 5, 2020, the standardized case reporting form was revised. Implementation of case reporting using this new form is ongoing among U.S. states and territories.

## Dataset Versions and Release Schedule

### Data are Considered Provisional

- The COVID-19 case surveillance data are dynamic; case reports can be modified at any time by the jurisdictions sharing COVID-19 data with CDC.
- CDC may update prior cases shared with CDC based on any updated information from jurisdictions.
- National case surveillance data are constantly changing. For instance, as new information is gathered about previously reported cases, health departments provide updated data to CDC. As more information and data become available, analyses might find changes in surveillance data and trends during a previously reported time window. Data may also be shared late with CDC due to the volume of COVID-19 cases.
- Annual finalized data: To create the final NNDSS data used in the annual tables, CDC works carefully with the reporting jurisdictions to reconcile the data received during the year until each state or territorial epidemiologist confirms that the data from their area are correct.

Version updates to the restricted and public datasets will be available every four weeks. The datasets will include all cases with an initial report date of case to CDC at least 14 days prior to the creation of the previously updated datasets. This month lag will allow adjustments to case reporting and ensure that time-dependent outcome data, including death, are accurately captured. Releases will be managed through github.com and will contain most recent and previous versions.

CDC's Case Surveillance Section routinely performs data quality assurance procedures (i.e., ongoing corrections and logic checks to address data errors). To date, the following data cleaning steps have been implemented:

- Questions that have been left unanswered (blank) on the case report form are re-classified to an *Unknown* value, if applicable to the question. For example, in the question "Was the patient hospitalized?", where the possible answer choices include "Yes", "No", or "Unknown", the missing value is re-coded to the *Unknown* answer option if the respondent did not answer the question.
- Logic checks are performed for date data. If an illogical date has been provided, CDC reviews the data with the reporting jurisdiction. For example, if a symptom onset date that is in the future is reported to CDC, this value is set to null until the reporting jurisdiction updates this information appropriately.
- The initial report date of the case to CDC is intended to be completed by the reporting jurisdiction when data are submitted. If blank, this variable is completed using the date the data file was first submitted to CDC.
- Additional data quality processing to recode free text data are ongoing. Data on symptoms, race and ethnicity, and healthcare worker status have been prioritized.

## Data Suppression

To prevent release of data that could be used to identify persons, data cells are suppressed for low frequency (<5) records. Records are never removed from the dataset, but individual field values are suppressed for geographic areas with low reporting counts or rare combinations of demographic characteristics (sex, age group, race/ethnicity). Suppressed values are re-coded to the *NA* answer option.

## Dataset Limitations

The COVID-19 case surveillance system is passive; data underestimate the true numbers of cases because of underdiagnosis or underreporting. Completeness of reporting is influenced by many factors (e.g., availability of diagnostic testing, resources and priorities health officials). Because reporting to CDC is voluntary, reporting practices vary by state and also depend on a variety of factors. Differences could exist between state-specific databases and CDC's COVID-19 surveillance database, though efforts are made to align CDC's database with state-specific data.

Although the case report form captures several outcomes, including hospitalization, ICU admission, and death, these data may be incomplete because outcomes are not yet known at the time of reporting (i.e., outcomes coded as *Unknown*). These data elements also may not represent final outcomes, as a patient's condition may have changed after case data submission but the case report was not updated.

## Data Requests from Agencies, Institutions, or Persons Outside the COVID-19 CDC EOC Response, Including Other CDC Employees

There will be no release of data in formats other than those described above, unless the format is more restrictive than described above. Requests for data must be made using the form, *Registration Information and Data Use Restrictions Agreement (RIDURA)*.

## Additional COVID-19 Data

COVID-19 data will be made available to the public as summary or aggregate count files, including total counts of cases and deaths by state and by county. These and other data on COVID-19 are available from multiple public locations:

https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html

https://www.cdc.gov/covid-data-tracker/index.html

https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html

https://www.cdc.gov/coronavirus/2019-ncov/php/open-america/surveillance-data-analytics.html